Guided project report

Dataset : scRNA-seq data from Drosophila melanogaster

Baptiste Rivoirard

UE: Advanced system biology (AdSP)

Introduction

Single-cell RNA sequencing (scRNA-seq) involves sequencing the RNA within individual cells, enabling the identification of distinct cell types and the study of cellular differentiation trajectories. While analysis methods are not yet fully standardized, the main steps of the pipeline include quality control and data filtering, normalization, selection of Highly Variable Genes (HVG), dimensionality reduction (PCA), clustering, and functional annotation.

Materials and Methods

We analyzed an unpublished scRNA-seq dataset from Drosophila melanogaster. This dataset includes a folder named rep2 containing the files barcode.tsv, features.tsv, and matrix.mtx. The analysis was performed in R using the Seurat package. The data was loaded as a sparse matrix with dimensions 17,753 by 2,331,657. After creating the Seurat object, we obtained 43,688 cells and 11,752 genes.

Results

The methods used, the parameters selected, and the results obtained are detailed in this section.

Once the Seurat object was created, the first step of the analysis pipeline involved basic preprocessing to filter out low-quality cells, such as under-sequenced cells, debris (broken cells or free-floating RNA fragments), and doublets/multiplets (multiple cells captured in a single droplet). We began with a descriptive analysis of the data using the summary function, which showed that the number of UMIs per cell had a median of 902 and an average of 3,900. For the number of genes per cell, the median was 486, and the average was 732.1. Next, we generated violin plots to visualize the distribution of UMIs and genes per cell, which guided the selection of thresholds for filtering low-quality cells (Figure 1A and 1B).

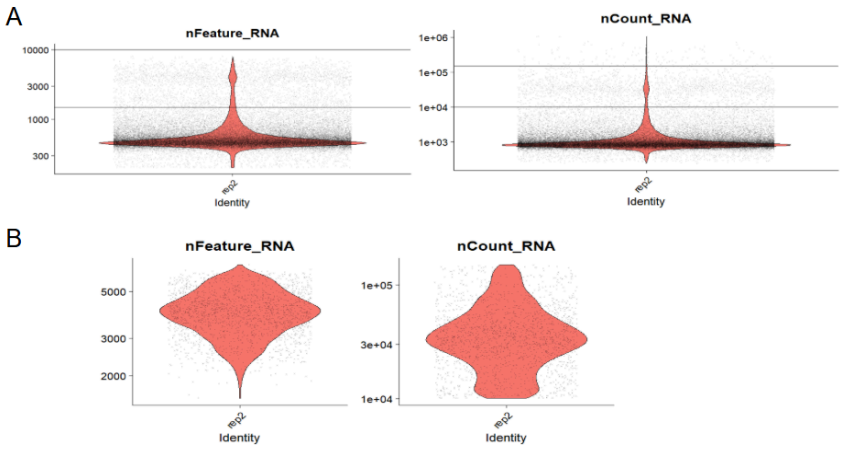

Figure 1: Cell filtering based on the number of UMI and the number of expressed genes

A) *Violin plot distribution of the number of UMIs and the number of expressed genes before filtering. The chosen thresholds are represented by the black lines.*
B) *Violin plot distribution of the number of UMIs and the number of expressed genes after filtering*

Based on the observed distributions, we chose the following thresholds: minimum UMIs: 10,000, maximum UMIs: 150,000, minimum Genes: 1,500, maximum Genes: 10,000 (Figure 1A). We then applied the thresholds using the subset function, retaining only cells with UMI and gene counts within these ranges (Figure 1B).

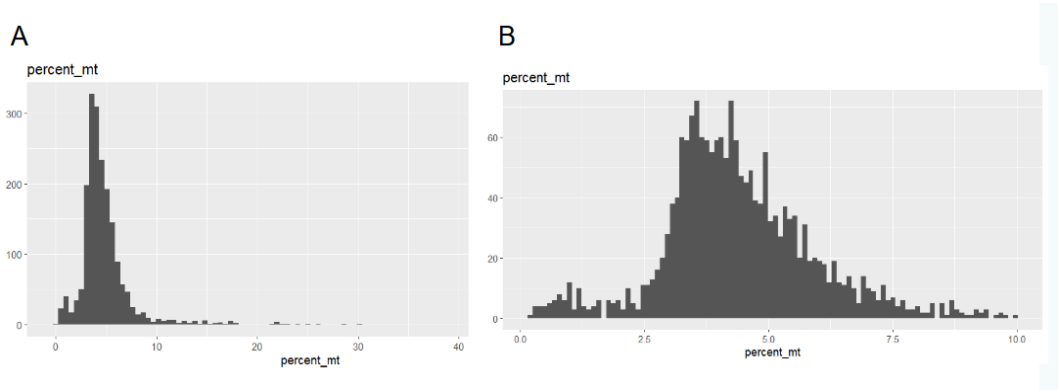

Figure 2: Mitochondrial RNA percentage in cells

A) *Histogram of mitochondrial RNA percentage in cells before filtering.*
B) *Histogram of mitochondrial RNA percentage in cells after filtering.*

Next, cells with a high percentage of mitochondrial RNA were filtered out, as this is indicative of dead or stressed cells, which are generally of poor quality. To achieve this, we retrieved mitochondrial genes specific to Drosophila melanogaster from the BioMart database. This step allowed us to identify the mitochondrial gene pattern for this species: “mt:” (indicating that mitochondrial genes in Drosophila start with this prefix). Using this pattern, we calculated the percentage of mitochondrial genes in each cell with the PercentageFeatureSet function. The average mitochondrial RNA percentage was 4.8%, which is excellent and suggests that our earlier filtering primarily retained high-quality cells (Figure 2A). However, some cells exhibited up to 50% mitochondrial RNA, necessitating additional filtering based on this parameter. We set a maximum mitochondrial RNA threshold of 10%, a commonly used cutoff that, in our case, retains the majority of cells while removing low-quality ones (Figure 2B). This resulted in a dataset composed solely of cells with low mitochondrial RNA content, ensuring a higher-quality dataset for downstream analyses.

Next, we filtered out doublets and multiplets present in our cell population. To achieve this, we used the scDblFinder package. The workflow began with converting the Seurat object into a SingleCellExperiment object. The algorithm then generated 1,500 artificial doublets and applied a k-nearest neighbors (kNN) model to identify and classify cells in our dataset that are likely doublets.The analysis detected 5.1% doublets, equating to 93 doublets among the 1,741 singlets. We proceeded with a final round of data filtering to remove these identified doublets, using the subset method. This step ensured a dataset composed exclusively of singlets, further improving the quality and reliability of subsequent analyses.

We then proceeded to data normalization, a critical step to account for cell-specific sequencing biases. Some cells express less RNA naturally, while others may appear under-sequenced due to technical artifacts. Normalization corrects for these discrepancies. We used Seurat's modern SCTransform method, which leverages generalized linear models (GLMs). In this case, we included regressions for nCount_RNA and percent_mt, which, while not strictly necessary, help minimize the influence of cells with very high RNA counts (potential doublets) or elevated mitochondrial content (potentially dead or stressed cells). After normalization, SCTransform selected 2,000 highly variable genes (HVGs), a standard threshold providing reliable results for clustering and population identification. SCTransform also performed scaling, where variable gene expression is centered around its mean with unit standard deviation. This ensures all genes contribute equally during dimensionality reduction and prevents highly expressed genes from dominating. The integration of these steps into SCTransform makes it both efficient and effective.

With these steps completed, we proceeded to Principal Component Analysis (PCA). While we initially selected 2,000 highly variable genes (HVGs), this number is still too high for downstream analyses, necessitating dimensionality reduction. We performed PCA using the RunPCA function on our filtered, normalized, and scaled Seurat object. (Refer to the annex for heatmaps of the most significant genes per PC.) We then examined the gene ontology (GO) of the top genes contributing to the first three principal components (PCs), which capture the largest variance. PC1: Genes primarily associated with cuticle development, muscle structure development, and neurogenesis regulation, along with other developmental biological processes (Figure 3). PC2: Genes involved in neuron differentiation, development, and morphogenesis. PC3: Genes tied to the development and morphogenesis of various organs and structures, such as the cuticle, appendages, imaginal discs, and antennae (Figure 3).

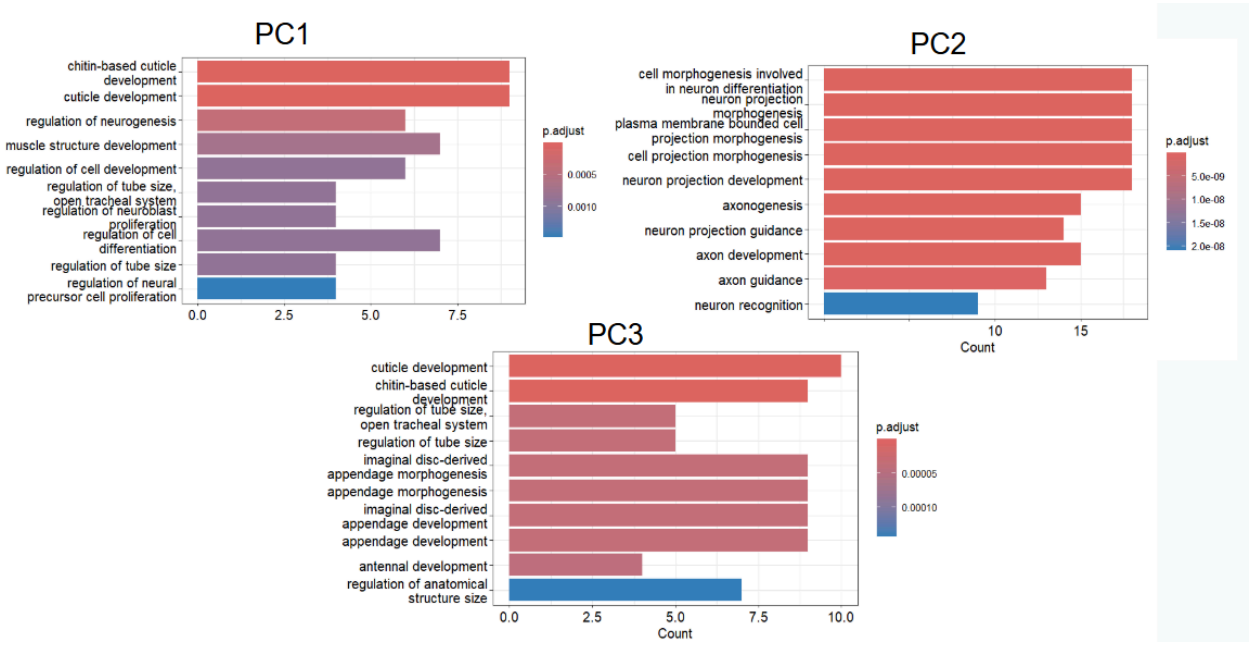

Figure 3: Gene ontology results for genes contributing the most to the first three principal components

After performing PCA, we will select the number of principal components to retain in order to preserve the maximum variance.

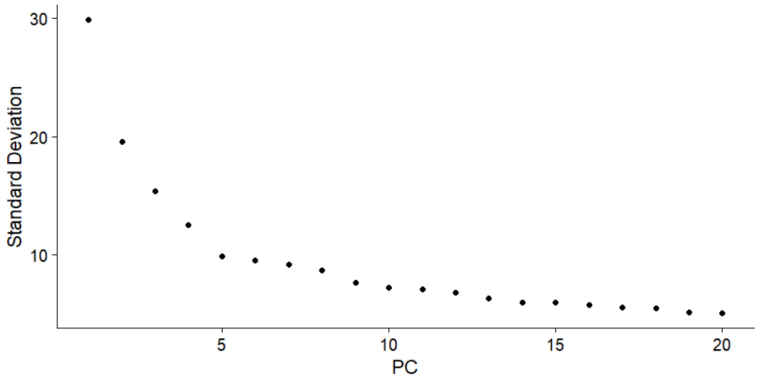

Figure 4: Proportion of the standard deviation explained by the first twenty principal components

The graph above shows the contribution of each principal component (Figure 4). Based on this graph and after testing different numbers of PCs, we chose to retain 15 PCs. This allows us to capture the majority of the variance and achieve the best clustering (testing with more PCs did not improve the clustering).

We are now ready to perform cell clustering. First, we create the s-NN graph (shared nearest neighbor graph) using the FindNeighbors method, then partition it with the FindClusters method, which is based on the Louvain algorithm. Several resolutions were tested for the latter step, and a resolution of 0.2 was selected as it produced a number of clusters consistent with the expected number of cell types.

Once the clusters are defined, they are visualized on a UMAP using the RunUMAP and DimPlot methods, resulting in 8 distinct clusters (Figure 5).

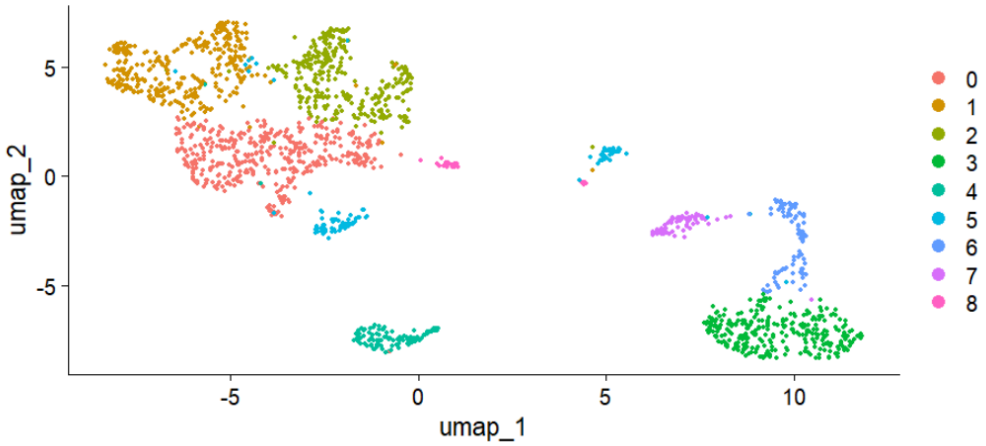

Figure 5: UMAP obtained after clustering

Using a list of marker genes for different cell types, we can now manually annotate each cluster from our UMAP as a specific cell type by examining the expression of these marker genes in the

clusters. This is done through violin plots showing the expression of each gene in the list across the different clusters, alongside an aggregate score calculated with the AddModuleScore function (Feature plots for the scores are shown in the annex). When looking at the expression of muscle-specific genes, several clusters contain cells expressing these genes (Figure 6A). However, when considering the overall expression of all the muscle-related genes, cluster 3 stands out as the cluster with cells expressing all muscle genes, showing the highest score for muscle cell type (Figure 6A). Some genes, such as blow, are expressed uniformly across the clusters. This is not surprising, as blow is involved not only in myoblast fusion but also in actin organization, making it present in other cell types beyond muscle cells (Figure 6A).

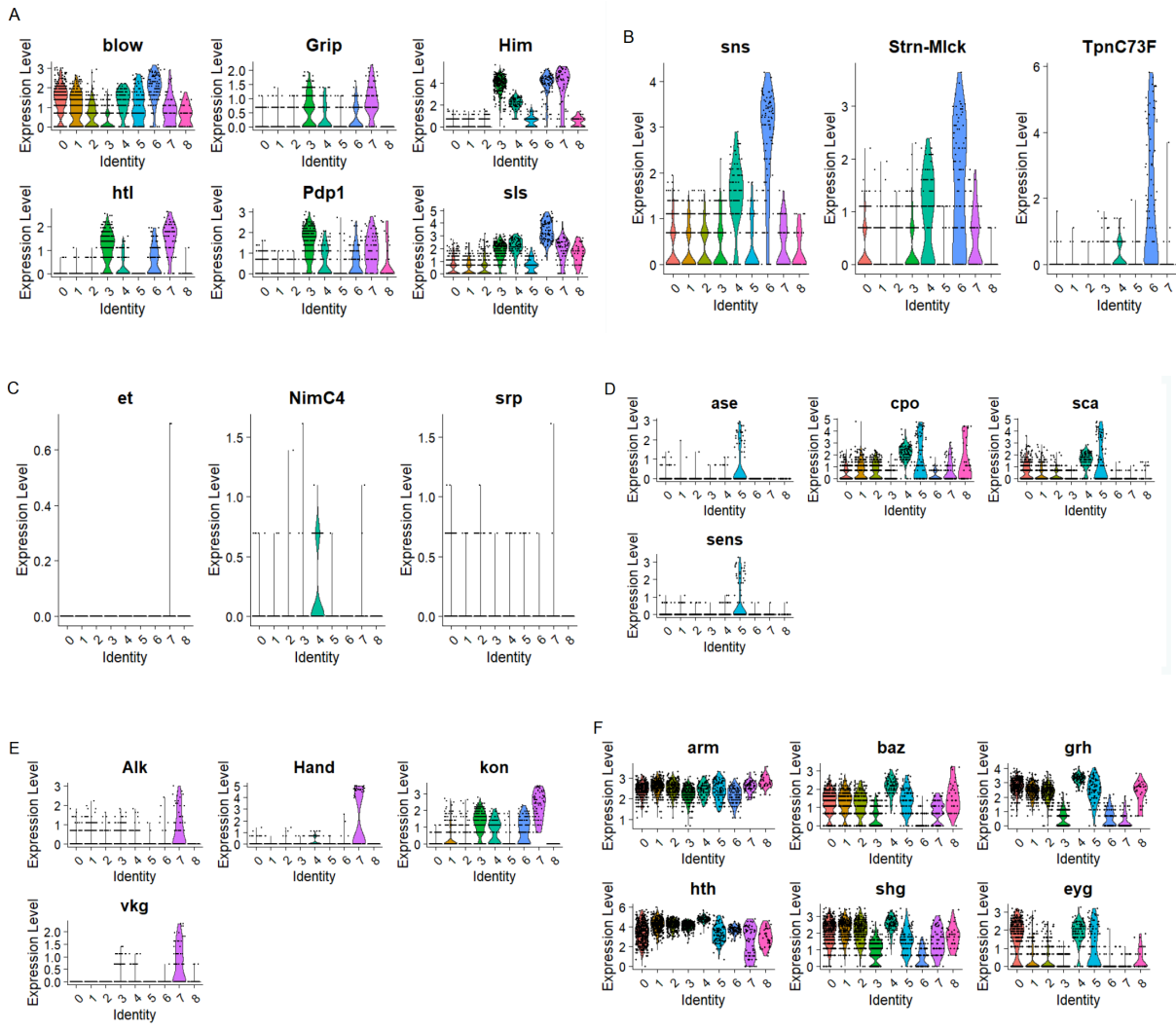

Figure 6: Violin plot of marker gene expression for cell types across different clusters

- *A) Expression of muscle marker genes*
- *B) Expression of muscle precursor marker genes*
- *C) Expression of hemocyte marker genes*
- *D) Expression of marker genes for precursor cells of external sensory organs*
- *E) Expression of tendon marker genes*
- *F) Expression of epithelial marker genes*

When focusing on the genes specific to muscle precursors, we observe that they are mainly expressed in cells from cluster 6 (except for the gene sis, which was not found in the dataset) (Figure 6B). The feature plot also shows that cells in cluster 6 have the highest score for this cell type. We observe some expression heterogeneity within the cells of this cluster, which can be explained by the fact that muscle precursor cells are in the process of developing into muscle cells, leading to significant variability within this cell type (Figure 6B).

The expression of genes characteristic of hemocytes is very low in all clusters, and only the expression of NimC4 was found in cluster 4, at a relatively low level. Therefore, if hemocytes are present in any cluster, it is likely to be cluster 4 (Figure 6C). It is also possible that only a small subset of cluster 4 consists of mature hemocytes, while some cells in this cluster may still be in a differentiation stage where they express little or none of the genes from this list. Cells in cluster 4 express genes from other lists as well, which suggests they could be part of a muscle population. However, this could also be explained by the fact that both muscle and hemocyte cells originate from mesodermal differentiation, so cells in cluster 4 may be in the process of differentiating into hemocytes while still expressing genes shared with muscle cells.

Regarding the expression of genes associated with precursors of external sensory organs, cells from cluster 5 emerge as the best candidates for this cell type. Indeed, cells in cluster 5 exhibit expression of all the genes on the list, unlike the other clusters (Figure 6D).

The expression of tendon-associated genes is almost exclusively found in cluster 7 (3 out of 4 genes). Although clusters 3 and 6 express one tendon gene, this can be explained by the tissue proximity between these cell types (tendon and muscle/pre-muscle) (Figure 6E). Furthermore, cells in cluster 7 have the highest score for the tendon cell type (see annex).

Finally, the genes associated with epithelial tissue show varying patterns of expression. Some are broadly expressed across all clusters, while others are not (Figure 6F). Overall, clusters 0, 1, and 2 express all the epithelial genes, and when the resolution is reduced, they are clustered together. Additionally, they do not strongly express genes from other cell types. Cells in these clusters also show a higher score for the epithelial cell type (see annex). Therefore, they are annotated as epithelial cells.

This results in a new UMAP with annotations for the cell types (Figure 7).

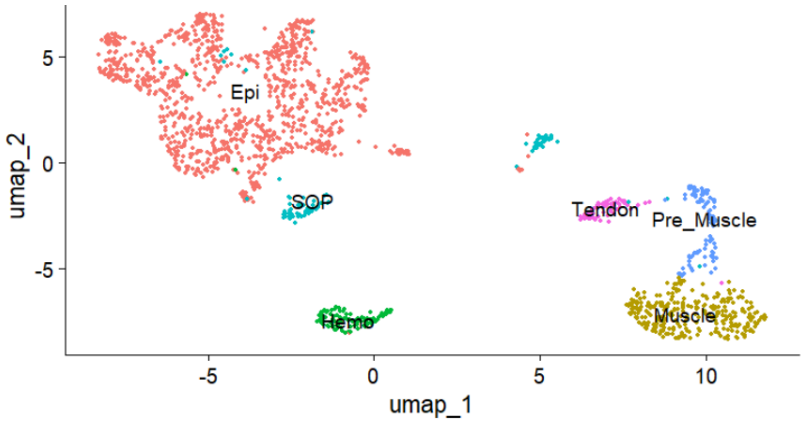

Figure 7: UMAP obtained after annotation

Finally, the last step of the analysis is to identify the marker genes for each cluster. To do this, we use the FindMarkers function, which identifies genes that are highly upregulated in a specific cluster compared to others (for the complete table, see Table 1 in the annex).

For clusters 0, 1, and 2, we mainly find genes that are either poorly categorized or genes such as Lcp65Ag2, CG17738, Cpr78E, and Cpr47Ec, which are associated with the cuticle and are closely linked to epithelial tissue, supporting our annotation.

For cluster 3, we find the gene CG7841, which is known to be involved in muscle cells. For cluster 4, we find the gene NimC4, which is part of the hemocyte marker gene list, further supporting our annotation. The marker genes for cluster 5 are particularly interesting, as two of the top three genes in this cluster, ase and sens, are part of the SOP gene list, strongly confirming our annotation. For cluster 6, we find the gene TpnC73F, which supports our annotation. In cluster 7, the gene hand is found among the markers, which was also part of the marker list for muscle precursor cells. Finally, for cluster 8, we find genes that are either uncharacterized or have no direct connection to our cellular type annotation, which does not support our annotation for this cluster.

Discussion

Several approaches for cell selection and filtering were tested for the analysis of this dataset. Two distinct cell populations are observed in the violin plots during the filtering of low-quality cells. After completing the analysis while retaining cells with a low number of UMIs and relatively few expressed genes (the majority of cells), we obtained poor clustering results. These cells formed a single large cluster, masking the heterogeneity of cellular populations, making the analysis impossible. Additionally, by keeping these cells, we observed an average mitochondrial RNA percentage of around 30%, which is very high and indicative of poor-quality cells (e.g., stressed or dead cells). Therefore, despite these cells representing the majority of the dataset, they were not included in the final analysis. It's possible that an issue during the experiment led to the presence of low-quality cells, or that a low sequencing depth (with fewer UMIs compared to the other population) was the cause, which could explain why this dataset was not published.

Regarding the marker gene lists for cell types, we observed that some genes were missing in the final analyzed cells (such as sis, eater, and zip). Additionally, the expression of certain genes was present in the majority, or even in all clusters, making them less useful for annotating cellular populations. However, it is normal for some cell types to express genes that are considered markers for another type. For example, cells in clusters 6 and 7 highly express genes from the muscle cell list, even though they are annotated as muscle precursor and tendon cells, respectively. This is not surprising and can be explained by the similarity of the tissues (muscle precursor and tendon cells share many genes with muscle cells).

Annexes

avg_log2FC	cluster	gene
4.05	0	Lcp65Ag2
3.67	0	CG10311
3.39	0	CG14752
5.06	1	Cpr78E
3.78	1	CG17738
3.77	1	grn
3.86	2	hh
3.76	2	Cpr47Ec
3.42	2	Doc1
6.66	3	CG15529
5.02	3	net
4.80	3	CG7841
4.66	4	CG4213
4.38	4	CG5399
4.33	4	NimC4
6.57	5	meru
6.44	5	ase
6.40	5	sens
8.70	6	TpnC73F
8.46	6	Hsc70-1
8.06	6	asRNA:Eig63F-2
9.75	7	Hand
9.70	7	eve
7.25	7	tin
9.37	8	CG4133
9.27	8	CG18765
8.72	8	CG5653

Table 1: *Marker genes for the different clusters*

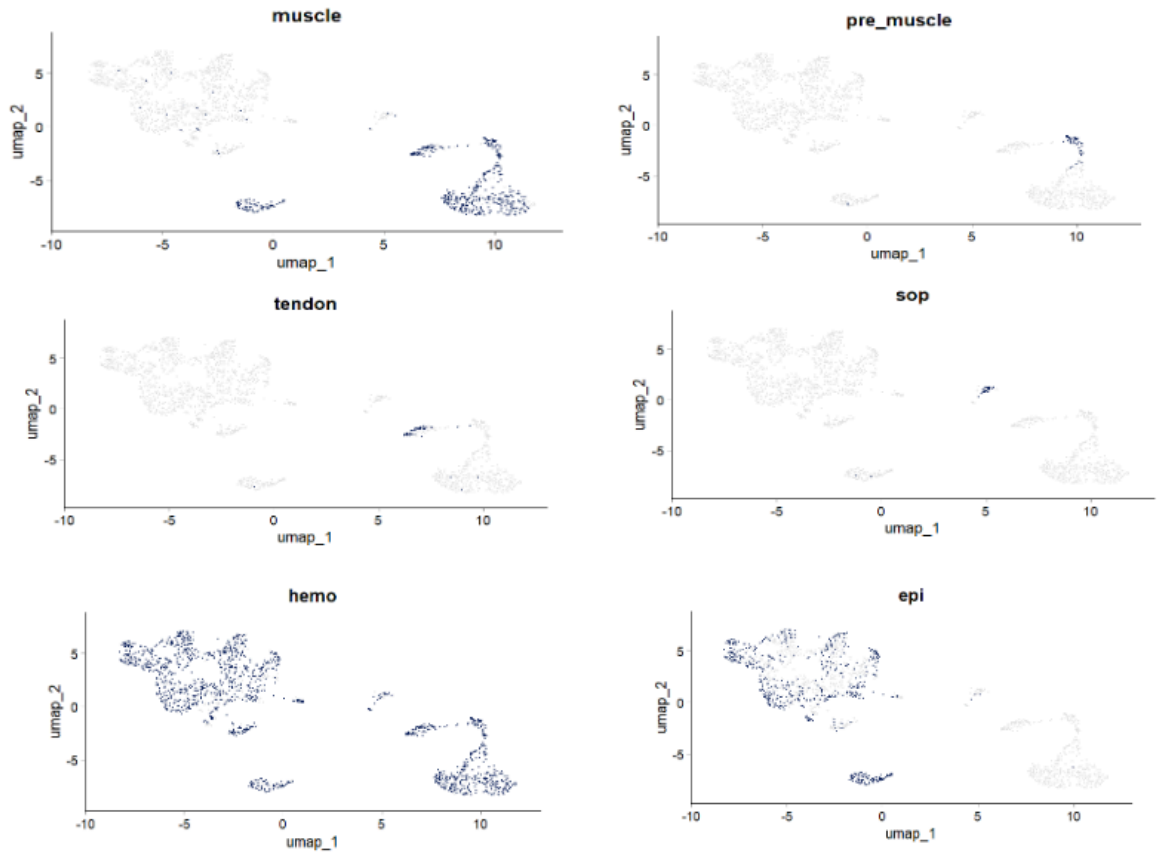

Figure Sup1: *FeaturePlot of cells annotated to different cell types based on their module score using AddModuleScore*

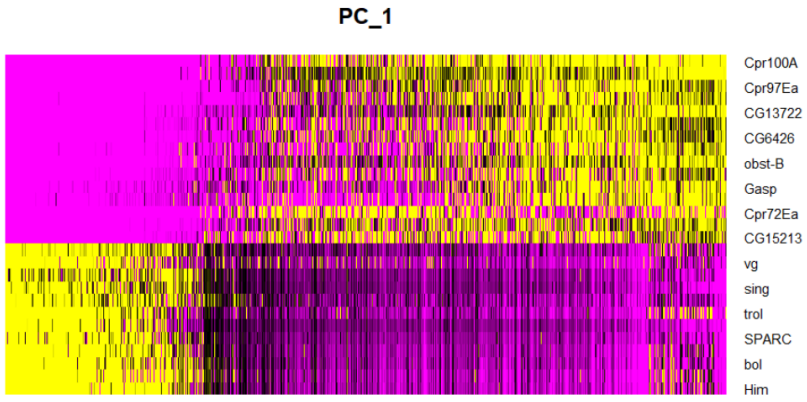

Figure Sup2: *Heatmap of genes most contributing to principal component 1*